Cell Genomics

CellPress
OPEN ACCESS

Perspective

Universal paralogs provide a window into evolution before the last universal common ancestor

Aaron D. Goldman,[1,2,*] Gregory P. Fournier,[3,*] and Betül Kaçar[4,*]
[1]Department of Biology, Oberlin College and Conservatory, Oberlin, OH, USA
[2]Blue Marble Space Institute of Science, Seattle, WA, USA
[3]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[4]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA
*Correspondence: agoldman@oberlin.edu (A.D.G.), g4nier@mit.edu (G.P.F.), bkacar@wisc.edu (B.K.)
https://doi.org/10.1016/j.xgen.2026.101140

SUMMARY

The study of early evolutionary history provides an account of how the foundational features of life as we know it first emerged. Phylogenetic analysis is a powerful method in the study of early evolution because it uses molecular evidence that has been inherited from the ancient organisms themselves. Here, we describe an important yet understudied type of protein family, universal paralogs, that retain phylogenetic signals from evolutionary events predating the last universal common ancestor of life, offering a unique window into early evolution. We survey recent advances in the study of universal paralogs and discuss how emerging computational tools enhance our ability to use these protein families to describe the very earliest stages of evolution with increasing detail and accuracy. Such research will greatly improve our understanding of how life emerged and subsequently evolved on the ancient Earth.

THE LAST UNIVERSAL COMMON ANCESTOR AND ITS GENOME

All extant life on Earth is descended from a last universal common ancestor (LUCA) that lived roughly four billion years ago.[1,2] The LUCA is the oldest organism that can be accessed through phylogenetic inference.[3,4] Over the past 25 years, numerous studies have sought to reconstruct the LUCA's genome or proteome.[2,5–9] Although these reconstructions differ in the exact set of specific genes or protein families they assign to the LUCA,[10,11] they consistently depict it as a relatively complex cellular organism.[11–13]

Conserved protein families involved in membrane transport and membrane bioenergetics suggest that the LUCA was a cellular organism with a functional membrane.[14] The similar conservation of RNA polymerase, multiple DNA repair polymerases, and other DNA-processing enzymes further suggests that it possessed a DNA genome,[15–17] although the lack of a universal replicative DNA polymerase enzyme has led some researchers to argue that the LUCA had an RNA-based genome[18,19] or a DNA/RNA hybrid genome.[20] The LUCA also appears to have encoded a nearly complete translation system similar to those seen across the biosphere today, including a complete ribosome,[21,22] the canonical genetic code,[23–25] and core translation-related proteins.[16,26]

Although the biosynthetic metabolism of the LUCA is less well understood, phylogenetic analyses have shown that it likely possessed enzymes that synthesize or interconvert nucleotides, amino acids, and sugars[27] and hint at a rudimentary carbon fixation pathway.[28] The energy metabolism of the LUCA is also poorly constrained by comparative genomics studies, although the presence of an ATP synthase[29,30] indicates that proton gradients were already being generated and used to synthesize ATP. Recent research also suggests that the LUCA may have run the anaerobic acetogenesis pathway in reverse as an energy metabolism.[2]

The LUCA, therefore, represents a cellular organism with a complete tripartite genetic system and complex metabolism. Evolutionary stages preceding the LUCA must have seen the development of nearly all of these core molecular systems that are foundational to cellular life. An accurate account of evolutionary history prior to the LUCA is, thus, paramount to our understanding of how life as we know it first took shape.

UNIVERSAL PARALOGS AND THEIR FUNCTIONS

While many strategies are used to study the origin and early evolution of life, from laboratory experiments in prebiotic chemistry to computer simulations of potential evolutionary processes, phylogenetic inference provides the most reliable way to directly study early evolutionary history using evidence preserved from the time when those ancient organisms lived. Though the LUCA represents the earliest type of organism that can be reconstructed through phylogenetic methods, some gene families retain evidence of evolutionary events prior to the LUCA. These gene families have at least two related versions, or paralogs, that can be reconstructed in the genome of the LUCA and remain present in most organisms' genomes across the tree of life today. Following recent literature, we refer to these protein families as "universal paralogs,"[2,31] although earlier literature

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

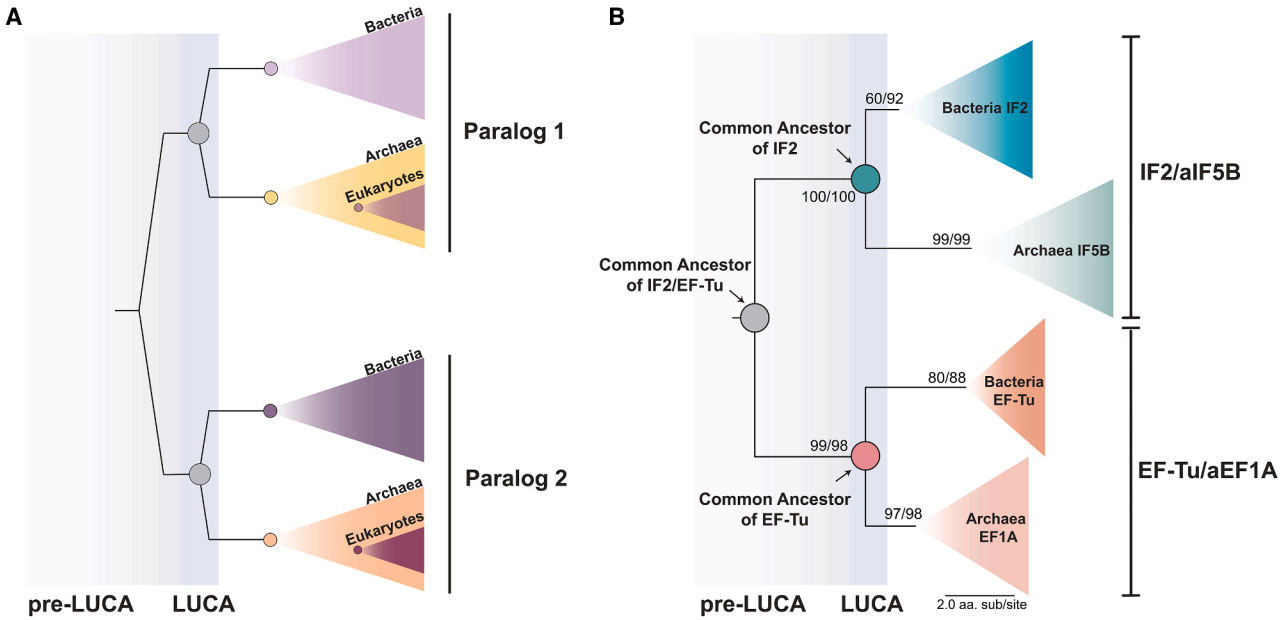

Figure 1. Phylogenetic topologies of universal paralog trees

(A) An idealized phylogeny of a universal paralog protein family in which two paralogs were present in the proteome of the LUCA. Each paralog generates a subtree tracing the species tree of life. The branch between the two sub-trees indicates a duplication event that must have occurred prior to the LUCA.

(B) An example of the universal paralog family containing initiation factors and elongation factors (see Table 1) showing branch support values ≥50%. The phylogeny was redrawn from data produced by Fer et al.[45] using FigTree, and clades were collapsed based on bacterial or archaeal proteins within each paralog subtree.

uses the term "anciently duplicated gene families" for the same phenomenon.[32,33]

Paralogy is a common phenomenon in evolution and a major source of new protein functions.[34,35] Paralogs are thought to originate primarily from gene duplication, although horizontal gene transfer may lead to a similar relationship between protein families called "xenologs."[36] Xenologs may be indistinguishable from paralogs in sufficiently old protein families. While most duplicate genes are subsequently lost as pseudogenes, those that are retained and fixed within lineages often evolve to perform a new function in a process called "neofunctionalization."[37,38] Sometimes one or both paralogs may evolve to perform only a component of the ancestral function in a process called "subfunctionalization."[39,40]

Paralogy is widespread in the genomes of organisms across the tree of life. Roughly 70% of protein-coding genes in the human genome are related to another protein-coding gene in the same genome.[41] The frequency of paralogy in prokaryotes is lower but still substantial.[42] For example, an estimated 46% of genes in the *E. coli* genome belong to paralogous families,[43] while about 17% of genes in the genome of the archaeon *Archaeoglobus fulgidus* belong to paralogous families.[44] The genome of the LUCA was likely no exception to this trend.

LUCA paralogs and their histories can be reconstructed with phylogenetic methods. In theory, if two paralogous proteins were encoded in the genome of the LUCA, then phylogenetic reconstruction of each individual paralog should separately produce a protein tree that traces the tree of life. When the entire family is represented on a single phylogeny, there would be two nodes representing the LUCA (one for each paralog) and a branch that connects them, tracing back to the pre-LUCA duplication event that produced the two paralogs from a single common ancestor (Figure 1). Because phylogenies of universal paralogs contain at least one pre-LUCA branch, they offer a unique opportunity to reconstruct events in molecular evolution that predate the LUCA.

Despite the prevalence of paralogy in extant genomes, only a small number of universal paralog protein families are currently known (Table 1). This paucity of known universal paralogs is not necessarily due to a lack of paralogs in the actual proteome of the LUCA. More likely, many of the paralogs that were present in the LUCA proteome cannot be clearly inferred, due to subsequent evolutionary events, including gene losses in descendant lineages, high levels of sequence divergence in a protein family, or extensive horizontal gene transfers that obscure the species tree signal. Lineage-specific gene losses can obscure the ancient origin of a protein family, as reconciliation methods require deep and robust phylogenetic signal to infer ancestral presence. High levels of sequence divergence due to rapid rates of evolution can lead to levels of sequence similarity in extant proteins that are below the statistical threshold for determining homology. Horizontally transferred genes can lead to the loss of the vertically inherited gene, a process called non-orthologous gene displacement,[46] which can also obscure the ancient origin of the protein family. For these reasons, it is likely that most protein families that were present in the LUCA may not be detectable through phylogenetic analyses.

CellPress
OPEN ACCESS

Table 1. Known universal paralog protein families and their molecular functions

Cellular function	Universal paralog protein family	Molecular function(s)
Amino acid metabolism	aminotransferases/carbamoyl transferases[47]	transfer of amine groups between different amino acids or between amino acids and amino acid-like compounds
	imidazole-4-carboxamide isomerase (HisA) and imidazole glycerol phosphate synthase (HisF)[48,49]	histidine biosynthesis
	carbamoyl phosphate synthetases[50]	transfer of carbamoyl phosphate compounds to amino acids
Translation process	aminoacyl-tRNA synthetases (several families)[51,52]	charging tRNA with their cognate amino acids.
	initiation factor IF2 and elongation factors EF-tu/EF1 and EF-G/EF2[45,53]	initiation of translation (IF2) and the regulation of the translation elongation cycle (EF-tu and EF-G)
Membrane translocation	signal recognition particle and its receptor[54]	recognizing membrane translocation signal peptides and initiating co-translation or post-translational transport across membranes
	ATP synthase catalytic and non-catalytic subunits[55]	ATP synthesis or ATP-driven proton transport
	ATP-binding cassette (ABC) transporters (multiple families)[33]	ATP-driven translocation of substrates across membranes

The known set of universal paralog families clusters around three main functional categories. Several universal paralog protein families are involved in the translation system, which is likely the most ancient molecular system retained in extant life today.[56,57] Other universal paralog families encode enzymes used in amino acid biosynthesis, which is related directly to translation and is similarly ancient.[24,58] The remaining known universal paralog families include proteins involved in the maintenance and function of biological membranes.[29,54] While the lack of a conserved phospholipid metabolism[59,60] has led some researchers to argue that the LUCA was not cellular,[61–63] proteins involved in the function of the membrane itself indicate that cellularity was indeed present at this time.[14]

PREVIOUS RESEARCH USING UNIVERSAL PARALOGS

Universal paralog protein families were first used as a tool to establish the species tree of life. In a typical phylogeny, the position of the root within any clade of organisms is determined by adding an outgroup to the tree, consisting of a set of more distantly related sequences. The node connecting the outgroup to the clade of interest indicates the location of the root of that clade. By definition, the universal tree of life has no taxonomic outgroup. However, phylogenies of universal paralog protein families effectively serve as one another's outgroups and so can be used to identify the location of the LUCA nodes (Figure 1), which indicate the location of the root of the universal tree.

Early studies rooted the tree of life in this manner, using, for example, the catalytic and non-catalytic subunits of ATP synthase,[29] aminoacyl tRNA synthetases,[51] and the signal recognition particle protein and its receptor.[54] These different universal paralog families indicated slightly different locations for the root of the universal species tree, which led Zhaxybayeva et al.[33] to conduct a broader study that identified 17 potential universal paralog protein families and compared the topologies of the species trees that they implied. The results of this study placed the root between the ancestors of the bacteria and the archaea/eukaryotes in most, but not all, trees. Some researchers at the time objected to this conclusion on the grounds that these universal paralog trees were subject to the artifact of long-branch attraction and therefore unreliable.[20,64–66]

Alternative rooting methods have since utilized amino acid usage biases in ribosomal proteins,[67] comparison of insertions and deletions in universal protein families,[68] and the analysis of different cellular properties,[69] all of which placed the root either within the bacteria or along the bacterial branch. More recently generated universal paralog trees containing proteins from bacteria, archaea, and eukaryotes[2,30,70] minimize the effects of long branch attraction through greater taxon sampling[71] and the more sophisticated evolutionary models used by current phylogenetic software.[72] These trees show the root between the bacteria and archaea, just as most universal paralog trees had done several decades ago.

Currently, universal paralog protein families are also emerging as a critical tool for reconstructing molecular biology prior to the time of the LUCA. The idea that universal paralogs could illuminate early evolutionary history was first proposed several decades ago,[73,74] around the same time that universal paralogs were also being used to root the tree of life. Today, modern methods of ancestral sequence reconstruction combined with protein structure and function prediction, as well as experimental approaches such as synthetic biology and ancient protein resurrection, have made direct investigation of pre-LUCA proteins possible.

One such line of research reconstructed the pre-LUCA ancestors of aminoacyl tRNA synthetases, and the frequencies of amino acids in their ancestral sequences were analyzed in order to investigate the late stages of genetic code evolution.[70,75] These enzymes are responsible for adding the correct amino acid to its cognate tRNA, thereby establishing the genetic code. In one study,[75] the composition of the common ancestor of leucyl-tRNA synthetase, isoleucyl-tRNA synthetase, and valyl-tRNA synthetase was shown to contain sites specific for leucine, isoleucine, and valine. In a second study,[70] the pre-LUCA ancestor of tryptophan tRNA synthetase and tyrosine tRNA synthetase was shown to be devoid of tryptophan amino

acids despite later versions of both synthetases containing tryptophan at several sites.

These results combined to illustrate important features of the later stages of genetic code evolution. The leucyl-tRNA, isoleucyl-tRNA, and valyl-tRNA synthetases appear to have evolved through subfunctionalization, as the ancestral protein sequence contains all three amino acids. This indicates that their ancestral protein did not discriminate between these three amino acids. This result suggested that the ability of early life forms to specifically incorporate these amino acids into genetically encoded proteins predates the evolution of the synthetases themselves, implying earlier, alternative aminoacylation processes. On the other hand, the reconstructed ancestral sequence of the tryptophanyl-tRNA and tyrosyl-tRNA synthetases has a statistically significant absence of tryptophan residues in its sequence, suggesting that the tryptophanyl-tRNA synthetase evolved through neofunctionalization. Taken together, these results support that the transition to a modern genetic code before the LUCA was a complex evolutionary process incorporating multiple mechanisms, including co-evolution with amino acid biosynthesis pathways.[76–78]

More recently, protein function prediction has been used to further characterize the pre-LUCA ancestors of universal paralog protein families. For example, the signal recognition particle and its receptor play an important role in the translocation of proteins into and across membranes. An analysis of this universal paralog family showed that the pre-LUCA ancestor was capable of several functions performed by either the signal recognition particle or its receptor, demonstrating that cellularity and complex membrane structure were well established prior to the time of the LUCA.[31] However, the signal recognition function itself, through which specific proteins are targeted to the membrane, could only be reconstructed for the LUCA node, not the pre-LUCA ancestral node, suggesting that this component of the system represents a neofunctionalization that occurred prior to the LUCA but after the initial duplication event.

Another study reconstructed the ancestor of the translation regulatory proteins, initiation factor IF2 and elongation factor EF-Tu.[45] IF2 is responsible for binding the initiator tRNA and facilitating the binding of the large and small ribosomal subunits around the mRNA, while EF-Tu controls the ratcheting forward of the translation process and also facilitates proofreading.[17] Sequence and structural comparisons of the tRNA binding site in the pre-LUCA IF2/EF-Tu ancestor to those of the modern versions of IF2 and EF-Tu showed that the pre-LUCA ancestor was likely capable of both functions, with a possible preference for initiator tRNA, and that IF2 and EF-Tu diverged through subfunctionalization before the emergence of the LUCA. The result that the LUCA translation factors were functional generalists is bolstered by kinetics experiments showing that resurrected ancestral EF-Tu proteins have equal preference for unrelated modern ribosomes, indicating the ribosomal core evolved before EF-Tu diversification.[79]

More recently, two studies further expanded the use of universal paralogs with new research strategies. One study[80] reconstructed the pre-LUCA ancestors of universally paralogous protein domains and compared their amino acid usages to post-LUCA protein domain reconstructions, showing trends that indicated the order in which specific amino acids were added to the genetic code. Another study[2] used phylogenies of five different universal paralogs to compare the dating of the emergence of the LUCA between the two LUCA nodes of each paralog family. While the accuracy and utility of these new methods remain to be assessed by follow-up studies, they suggest that universal paralogs will likely play an increasingly important role in the study of early life as new applications continue to be developed.

FUTURE RESEARCH ON PRE-LUCA EVOLUTION

The current list of known universal paralogs (Table 1) may only be a small subset of available protein families that exhibit a pre-LUCA duplication event. In early attempts to catalog universal paralog protein families,[32,33] taxon sampling may have been too sparse due to limited genomic data to identify all the protein families that have a strong signal of being universal paralog proteins. At the same time, earlier methods for phylogenetic reconstruction may have limited the ability to reliably trace histories to the time of the LUCA and its predecessors. Expanding the number of known universal paralog protein families or other protein families with evidence of a pre-LUCA evolutionary event will greatly enhance our understanding of early molecular evolution.

Current and emerging phylogenetics methods, along with the availability of sequence data from a broader range of microbial taxa, may lead to the discovery of new universal paralog protein families. Denser taxon sampling could provide evidence of rare universal paralogs that were lost in many post-LUCA lineages. Greater taxon sampling is also known to improve phylogenetic accuracy.[81,82] In parallel to the growth of taxonomically broad sequence data, several improvements in phylogenetic inference allow for more accurate recovery of deep evolutionary relationships, including more sophisticated models of evolutionary rates[83] and amino acid substitutions.[84] Because protein structure is usually more conserved than amino acid sequence over the course of evolution,[85] researchers have also used structural comparisons to identify very deep evolutionary relationships obscured by extreme protein sequence divergence.[86] The integration of ancestral sequence reconstruction with structure prediction now allows the assessment of structural variation across entire protein lineages.[87] The development of more accurate machine learning-based protein structure predictions may further improve the power of these methods, making extensive structure-based homology searches tractable.

As these methodological advances expand our ability to infer ancient relationships, they may also lead to the discovery of other kinds of protein families that predate the time of the LUCA. One recent study[2] estimated that the LUCA genome encoded ~2,657 proteins, but the same study only identified 399 specific protein families with a probability $\geq$ 0.75 of being present in the LUCA proteome. It is likely that this disparity is mostly due to gene loss following the LUCA, and some of these losses could obscure universal paralog histories (Figure 2B). An example of this sort of protein family is the rRNA adenine N(6)-methyltransferase,[88] which is not universally paralogous across genomes today but likely underwent a pre-LUCA duplication according to phylogenetic analysis. With increasingly

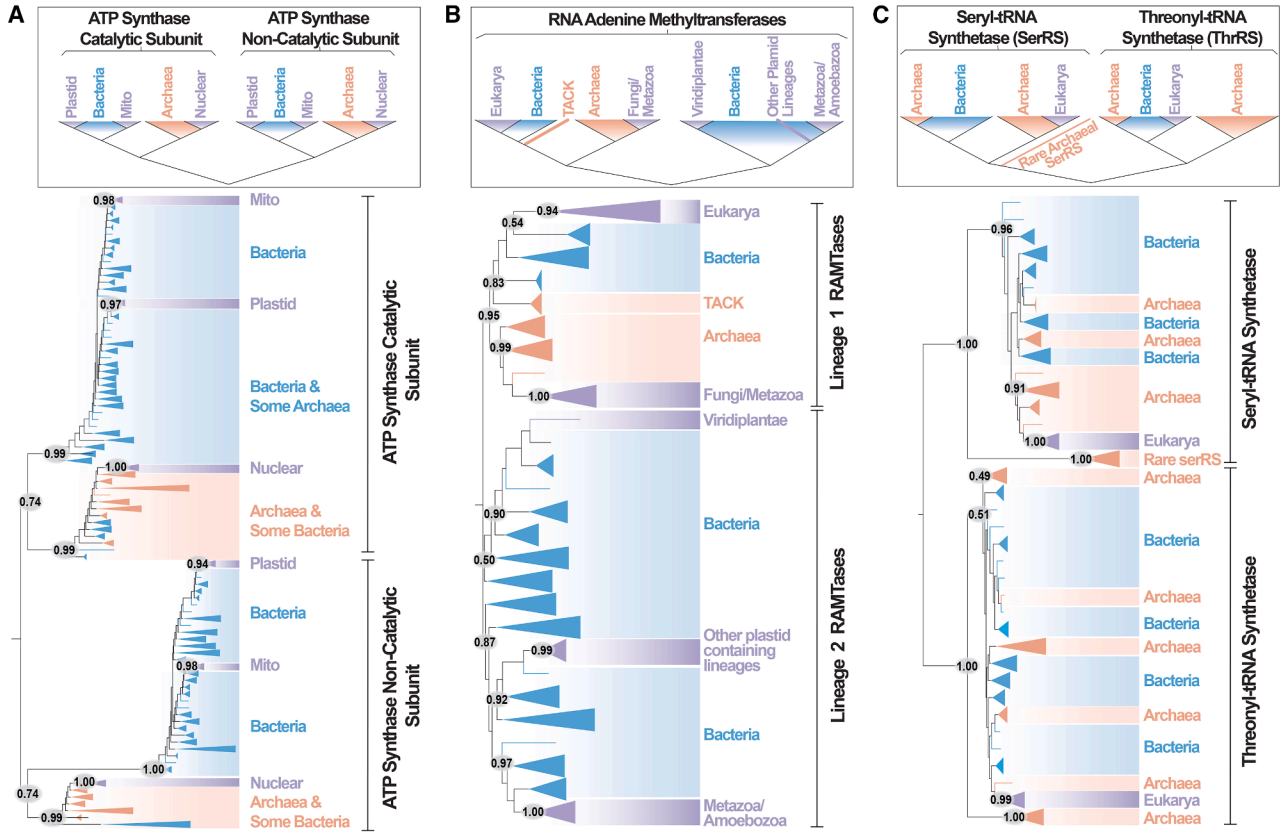

Figure 2. Examples of phylogenies indicating pre-LUCA evolutionary events

(A) The ATP synthase catalytic and non-catalytic subunits form a phylogenetic topology typical of universal paralog protein families with a LUCA node for both the catalytic subunit clade and non-catalytic subunit clade.[30]

(B) A family of RNA methyltransferase enzymes and mitochondrial transcription factors in which one large clade appears to have a LUCA node connected to a long branch with many bacterial paralogs. The most parsimonious explanation for this tree topology is a pre-LUCA duplication and an early loss of one paralog in the archaeal ancestral lineage.[88]

(C) The family of seryl-tRNA synthetase and threonyl-tRNA synthetase enzymes forms a typical universal paralog topology. However, a rare seryl-tRNA synthetase found in some euryarchaeota forms a branch between the two LUCA nodes on the tree, indicating a horizontal gene transfer from a pre-LUCA lineage that is no longer extant.[89] All trees were redrawn using FigTree, showing branch support values ≥50%. Clades were collapsed based on bacterial or archaeal proteins within each paralog subtree.

accurate gene tree/species tree reconciliation methods, it may be possible to detect a larger number of protein families that underwent pre-LUCA duplications but are no longer retained as paralogs across the entire tree of life today.

Relatedly, some protein families may show signs of ancient horizontal gene transfers that can also shed light on pre-LUCA evolution (Figure 2C). For example, some archaeal seryl-tRNA synthetases are separated from other archaeal and bacterial sequences by a very long branch, inconsistent with a straightforward interpretation of having a common ancestor mapping to the LUCA. Rather, this has been interpreted as the result of an ancient horizontal gene transfer from a pre-LUCA lineage into the archaeal domain. Similar patterns have been observed for pyrrolysyl-tRNA synthetase and possibly threonyl-tRNA synthetase.[89] While not paralogs in the strict sense, protein families undergoing pre-LUCA horizontal gene transfers from now-extinct lineages offer another resource for reconstructing common ancestry that predates the LUCA, and many more such instances may await detection with a focused effort.

In addition to expanding the set of known protein families with pre-LUCA phylogenetic branches, the ability to reconstruct the ancestral sequences and functions of pre-LUCA ancestral proteins is dramatically improving, a trend that is likely to accelerate in the next decade and beyond. Deep learning methods have revolutionized protein structure and function prediction over the last several years.[90] Similarly, emerging hardware technologies such as AI and quantum processor chips may make precise molecular dynamics modeling of complex biomolecular systems increasingly feasible in the near future.[91,92] These methods can be paired with ancestral sequence reconstructions of the LUCA and pre-LUCA nodes of universal paralog protein families to more accurately reconstruct details about the molecular structures, functions, and enzymatic mechanisms of the ancestors of universal paralogs.

Beyond understanding the historical process of early evolution, the ability to examine deep ancestors and reconstruct their protein sequences may provide new resources for bioengineering. By leveraging emerging computational tools alongside ancestral sequence reconstruction, researchers could access a catalog of proteins and their domains and motifs that were capable of carrying out fundamental catalytic roles but that no longer exist in the same form today, thereby expanding the known functional repertoire of life and providing novel catalysts for use in biotechnology and related fields.[93]

Universal paralog families are an important tool for understanding early evolution from a phylogenetic perspective, offering a unique and valuable form of evidence about molecular evolution prior to the LUCA. The phylogenetic study of ancient life is constrained by several fundamental limitations. Both gene loss across multiple lineages and low levels of conservation in some gene families can obscure the ancient origin of those gene families. Furthermore, in the absence of an extensive diagnostic fossil record, the dependence of molecular phylogenetics on conserved gene sequences means that periods of evolution that predated the emergence of the genetic system cannot be studied. Even so, emerging technologies across a number of areas of computational biology and synthetic biology will expand our ability to reconstruct pre-LUCA evolution using these protein families. As our understanding of the LUCA solidifies,[2,11] universal paralog protein families will provide an indispensable tool for pushing our understanding of early evolutionary history even further back in time, thereby describing the foundational processes that shaped life as we know it today.

ACKNOWLEDGMENTS

This work was supported by the National Aeronautics and Space Administration (NASA) Interdisciplinary Consortium for Astrobiology Research: Metal Utilization and Selection Across Eons (MUSE) 80NSSC17K0296 (B.K.); the NASA Exobiology Program NNH23ZDA001N as part of the NASA LIFE Research Coordination Network (B.K.); and the NASA Interdisciplinary Consortium for Astrobiology Research grant 80NSSC19M0069 (A.D.G.).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Betts, H.C., Puttick, M.N., Clark, J.W., Williams, T.A., Donoghue, P.C.J., and Pisani, D. (2018). Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nat. Ecol. Evol. *2*, 1556–1562. https://doi.org/10.1038/s41559-018-0644-x.
2. Moody, E.R.R., Alvarez-Carretero, S., Mahendrarajah, T.A., Clark, J.W., Betts, H.C., Dombrowski, N., Szantho, L.L., Boyle, R.A., Daines, S., Chen, X., et al. (2024). The nature of the last universal common ancestor and its impact on the early Earth system. Nat. Ecol. Evol. *8*, 1654–1666. https://doi.org/10.1038/s41559-024-02461-1.
3. Bell, E.A., Boehnke, P., Harrison, T.M., and Mao, W.L. (2015). Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. Proc. Natl. Acad. Sci. USA *112*, 14518–14521. https://doi.org/10.1073/pnas.1517557112.
4. Woese, C. (1998). The universal ancestor. Proc. Natl. Acad. Sci. USA *95*, 6854–6859. https://doi.org/10.1073/pnas.95.12.6854.
5. Harris, J.K., Kelley, S.T., Spiegelman, G.B., and Pace, N.R. (2003). The genetic core of the universal ancestor. Genome Res. *13*, 407–412. https://doi.org/10.1101/gr.652803.
6. Ranea, J.A.G., Sillero, A., Thornton, J.M., and Orengo, C.A. (2006). Protein superfamily evolution and the last universal common ancestor (LUCA). J. Mol. Evol. *63*, 513–525. https://doi.org/10.1007/s00239-005-0289-7.
7. Mirkin, B.G., Fenner, T.I., Galperin, M.Y., and Koonin, E.V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol. *3*, 2. https://doi.org/10.1186/1471-2148-3-2.
8. Wang, M., Yafremava, L.S., Caetano-Anollés, D., Mittenthal, J.E., and Caetano-Anollés, G. (2007). Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res. *17*, 1572–1585. https://doi.org/10.1101/gr.6454307.
9. Yang, S., Doolittle, R.F., and Bourne, P.E. (2005). Phylogeny determined by protein domain content. Proc. Natl. Acad. Sci. USA *102*, 373–378. https://doi.org/10.1073/pnas.0408810102.
10. Berkemer, S.J., and McGlynn, S.E. (2020). A New Analysis of Archaea-Bacteria Domain Separation: Variable Phylogenetic Distance and the Tempo of Early Evolution. Mol. Biol. Evol. *37*, 2332–2340. https://doi.org/10.1093/molbev/msaa089.
11. Crapitto, A.J., Campbell, A., Harris, A.J., and Goldman, A.D. (2022). A consensus view of the proteome of the last universal common ancestor. Ecol. Evol. *12*, e8930. https://doi.org/10.1002/ece3.8930.
12. Becerra, A., Delaye, L., Islas, S., and Lazcano, A. (2007). The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains. Annu. Rev. Ecol. Evol. Syst. *38*, 361–379. https://doi.org/10.1146/annurev.ecolsys.38.091206.095825.
13. Goldman, A.D., and Fournier, G.P. (2024). The very early evolution of biological complexity. Trends Genet. *40*, 912–913. https://doi.org/10.1016/j.tig.2024.09.001.
14. Kailing, F., Lieberman, J., Wang, J., Turner, J.L., and Goldman, A.D. (2024). Evolution of Cellular Organization Along the First Branches of the Tree of Life. J. Mol. Evol. *92*, 618–623. https://doi.org/10.1007/s00239-024-10188-7.
15. Goldman, A.D., and Landweber, L.F. (2012). Oxytricha as a modern analog of ancient genome evolution. Trends Genet. *28*, 382–388. https://doi.org/10.1016/j.tig.2012.03.010.
16. Koonin, E.V., Krupovic, M., Ishino, S., and Ishino, Y. (2020). The replication machinery of LUCA: common origin of DNA replication and transcription. BMC Biol. *18*, 61. https://doi.org/10.1186/s12915-020-00800-9.
17. Fer, E., Yao, T., McGrath, K.M., Goldman, A.D., and Kaçar, B. (2025). The origins and evolution of translation factors. Trends Genet. *41*, 590–600. https://doi.org/10.1016/j.tig.2025.02.004.
18. Forterre, P. (2006). Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. Proc. Natl. Acad. Sci. USA *103*, 3669–3674. https://doi.org/10.1073/pnas.0510333103.
19. Poole, A.M., and Logan, D.T. (2005). Modern mRNA proofreading and repair: clues that the last universal common ancestor possessed an RNA genome? Mol. Biol. Evol. *22*, 1444–1455. https://doi.org/10.1093/molbev/msi132.
20. Leipe, D.D., Aravind, L., and Koonin, E.V. (1999). Did DNA replication evolve twice independently? Nucleic Acids Res. *27*, 3389–3401. https://doi.org/10.1093/nar/27.17.3389.
21. Hoeppner, M.P., Gardner, P.P., and Poole, A.M. (2012). Comparative analysis of RNA families reveals distinct repertoires for each domain of life. PLoS Comput. Biol. *8*, e1002752. https://doi.org/10.1371/journal.pcbi.1002752.
22. Hsiao, C., Mohan, S., Kalahar, B.K., and Williams, L.D. (2009). Peeling the onion: ribosomes are ancient molecular fossils. Mol. Biol. Evol. *26*, 2415–2425. https://doi.org/10.1093/molbev/msp163.

23. Crick, F.H. (1968). The origin of the genetic code. J. Mol. Biol. *38*, 367–379. https://doi.org/10.1016/0022-2836(68)90392-6.

24. Knight, R.D., Freeland, S.J., and Landweber, L.F. (1999). Selection, history and chemistry: the three faces of the genetic code. Trends Biochem. Sci. *24*, 241–247. https://doi.org/10.1016/s0968-0004(99)01392-4.

25. Knight, R.D., Freeland, S.J., and Landweber, L.F. (2001). Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Genet. *2*, 49–58. https://doi.org/10.1038/35047500.

26. Goldman, A.D., Samudrala, R., and Baross, J.A. (2010). The evolution and functional repertoire of translation proteins following the origin of life. Biol. Direct *5*, 15. https://doi.org/10.1186/1745-6150-5-15.

27. Goldman, A.D., Baross, J.A., and Samudrala, R. (2012). The enzymatic and metabolic capabilities of early life. PLoS One *7*, e39912. https://doi.org/10.1371/journal.pone.0039912.

28. Braakman, R., and Smith, E. (2012). The emergence and early evolution of biological carbon-fixation. PLoS Comput. Biol. *8*, e1002455. https://doi.org/10.1371/journal.pcbi.1002455.

29. Gogarten, J.P., and Taiz, L. (1992). Evolution of proton pumping ATPases: Rooting the tree of life. Photosynth. Res. *33*, 137–146. https://doi.org/10.1007/BF00039176.

30. Mahendrarajah, T.A., Moody, E.R.R., Schrempf, D., Szánthó, L.L., Dombrowski, N., Davín, A.A., Pisani, D., Donoghue, P.C.J., Szöllősi, G.J., Williams, T.A., and Spang, A. (2023). ATP synthase evolution on a cross-braced dated tree of life. Nat. Commun. *14*, 7456. https://doi.org/10.1038/s41467-023-42924-w.

31. Harris, A.J., and Goldman, A.D. (2021). The very early evolution of protein translocation across membranes. PLoS Comput. Biol. *17*, ARTN e1008623. https://doi.org/10.1371/journal.pcbi.1008623.

32. Kollman, J.M., and Doolittle, R.F. (2000). Determining the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs. J. Mol. Evol. *51*, 173–181. https://doi.org/10.1007/s002390010078.

33. Zhaxybayeva, O., Lapierre, P., and Gogarten, J.P. (2005). Ancient gene duplications and the root(s) of the tree of life. Protoplasma *227*, 53–64. https://doi.org/10.1007/s00709-005-0135-1.

34. Zhang, J. (2003). Evolution by gene duplication: an update. Trends Ecol. Evol. *18*, 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8.

35. Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. *9*, 938–950. https://doi.org/10.1038/nrg2482.

36. Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. *39*, 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725.

37. Ono, S. (1973). Ancient linkage groups and frozen accidents. Nature *244*, 259–262. https://doi.org/10.1038/244259a0.

38. Soskine, M., and Tawfik, D.S. (2010). Mutational effects and the evolution of new protein functions. Nat. Rev. Genet. *11*, 572–582. https://doi.org/10.1038/nrg2808.

39. Hughes, A.L. (2005). Gene duplication and the origin of novel proteins. Proc. Natl. Acad. Sci. USA *102*, 8791–8792. https://doi.org/10.1073/pnas.0503922102.

40. Jensen, R.A. (1976). Enzyme recruitment in evolution of new function. Annu. Rev. Microbiol. *30*, 409–425. https://doi.org/10.1146/annurev.mi.30.100176.002205.

41. Ibn-Salem, J., Muro, E.M., and Andrade-Navarro, M.A. (2017). Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res. *45*, 81–91. https://doi.org/10.1093/nar/gkw813.

42. Gevers, D., Vandepoele, K., Simillon, C., and Van de Peer, Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. *12*, 148–154. https://doi.org/10.1016/j.tim.2004.02.007.

43. Koonin, E.V., Tatusov, R.L., and Rudd, K.E. (1995). Sequence similarity analysis of Escherichia coli proteins: functional and evolutionary implications. Proc. Natl. Acad. Sci. USA *92*, 11921–11925. https://doi.org/10.1073/pnas.92.25.11921.

44. von Jan, M., Lapidus, A., Del Rio, T.G., Copeland, A., Tice, H., Cheng, J.F., Lucas, S., Chen, F., Nolan, M., Goodwin, L., et al. (2010). Complete genome sequence of Archaeoglobus profundus type strain (AV18). Stand. Genomic Sci. *2*, 327–346. https://doi.org/10.4056/sigs.942153.

45. Fer, E., McGrath, K.M., Guy, L., Hockenberry, A.J., and Kaçar, B. (2022). Early divergence of translation initiation and elongation factors. Protein Sci. *31*, e4393. https://doi.org/10.1002/pro.4393.

46. Koonin, E.V., Mushegian, A.R., and Bork, P. (1996). Non-orthologous gene displacement. Trends Genet. *12*, 334–336.

47. Labedan, B., Boyen, A., Baetens, M., Charlier, D., Chen, P., Cunin, R., Durbeco, V., Glansdorff, N., Herve, G., Legrain, C., et al. (1999). The evolutionary history of carbamoyltransferases: A complex set of paralogous genes was already present in the last universal common ancestor. J. Mol. Evol. *49*, 461–473. https://doi.org/10.1007/pl00006569.

48. Charlebois, R.L., Sensen, C.W., Doolittle, W.F., and Brown, J.R. (1997). Evolutionary analysis of the hisCGABdFDEHI gene cluster from the archaeon Sulfolobus solfataricus P2. J. Bacteriol. *179*, 4429–4432. https://doi.org/10.1128/jb.179.13.4429-4432.1997.

49. Fani, R., Liò, P., Chiarelli, I., and Bazzicalupo, M. (1994). The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the hisA and hisF genes. J. Mol. Evol. *38*, 489–495. https://doi.org/10.1007/BF00178849.

50. Lawson, F.S., Charlebois, R.L., and Dillon, J.A. (1996). Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. Mol. Biol. Evol. *13*, 970–977. https://doi.org/10.1093/oxfordjournals.molbev.a025665.

51. Brown, J.R., and Doolittle, W.F. (1995). Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. USA *92*, 2441–2445. https://doi.org/10.1073/pnas.92.7.2441.

52. Brown, J.R., Robb, F.T., Weiss, R., and Doolittle, W.F. (1997). Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. J. Mol. Evol. *45*, 9–16. https://doi.org/10.1007/pl00006206.

53. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA *86*, 9355–9359. https://doi.org/10.1073/pnas.86.23.9355.

54. Gribaldo, S., and Cammarano, P. (1998). The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. J. Mol. Evol. *47*, 508–516. https://doi.org/10.1007/pl00006407.

55. Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T., et al. (1989). Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA *86*, 6661–6665. https://doi.org/10.1073/pnas.86.17.6661.

56. Fox, G.E. (2010). Origin and evolution of the ribosome. Cold Spring Harbor Perspect. Biol. *2*, a003483. https://doi.org/10.1101/cshperspect.a003483.

57. Freeland, S.J., Knight, R.D., and Landweber, L.F. (1999). Do proteins predate DNA? Science *286*, 690–692. https://doi.org/10.1126/science.286.5440.690.

58. Hernandez-Montes, G., Diaz-Mejia, J.J., Perez-Rueda, E., and Segovia, L. (2008). The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. Genome Biol. *9*, R95. https://doi.org/10.1186/gb-2008-9-6-r95.

59. Pereto, J., Lopez-Garcia, P., and Moreira, D. (2004). Ancestral lipid biosynthesis and early membrane evolution. Trends Biochem. Sci. *29*, 469–477. https://doi.org/10.1016/j.tibs.2004.07.002.

60. Sojo, V. (2019). Why the Lipid Divide? Membrane Proteins as Drivers of the Split between the Lipids of the Three Domains of Life. Bioessays *41*, e1800251. https://doi.org/10.1002/bies.201800251.

61. Koga, Y., Kyuragi, T., Nishihara, M., and Sone, N. (1998). Did archaeal and bacterial cells arise independently from noncellular Precursors? A hypothesis stating that the advent of membrane phospholipid with enantiomeric glycerophosphate backbones caused the separation of the two lines of descent. J. Mol. Evol. *47*, 631. https://doi.org/10.1007/pl00006419.

62. Martin, W., and Russell, M.J. (2003). On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. Philos. Trans. R. Soc. Lond. B Biol. Sci. *358*, 59–85. https://doi.org/10.1098/rstb.2002.1183.

63. Wachtershauser, G. (1988). Before enzymes and templates: theory of surface metabolism. Microbiol. Rev. *52*, 452–484. https://doi.org/10.1128/mr.52.4.452-484.1988.

64. Lopez, P., Forterre, P., and Philippe, H. (1999). The root of the tree of life in the light of the covarion model. J. Mol. Evol. *49*, 496–508. https://doi.org/10.1007/pl00006572.

65. Philippe, H., and Forterre, P. (1999). The rooting of the universal tree of life is not reliable. J. Mol. Evol. *49*, 509–523. https://doi.org/10.1007/pl00006573.

66. Forterre, P., and Philippe, H. (1999). Where is the root of the universal tree of life? Bioessays *21*, 871–879. https://doi.org/10.1002/(SICI)1521-1878(199910)21:10<871::AID-BIES10>3.0.CO;2-Q.

67. Fournier, G.P., and Gogarten, J.P. (2010). Rooting the ribosomal tree of life. Mol. Biol. Evol. *27*, 1792–1801. https://doi.org/10.1093/molbev/msq057.

68. Lake, J.A., Skophammer, R.G., Herbold, C.W., and Servin, J.A. (2009). Genome beginnings: rooting the tree of life. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 2177–2185. https://doi.org/10.1098/rstb.2009.0035.

69. Cavalier-Smith, T. (2006). Rooting the tree of life by transition analyses. Biol. Direct *1*, 19. https://doi.org/10.1186/1745-6150-1-19.

70. Fournier, G.P., and Alm, E.J. (2015). Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. J. Mol. Evol. *80*, 171–185. https://doi.org/10.1007/s00239-015-9672-1.

71. Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. *47*, 9–17. https://doi.org/10.1080/106351598260996.

72. Bergsten, J. (2005). A review of long-branch attraction. Cladistics *21*, 163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x.

73. Lazcano, A. (1995). Cellular evolution during the early Archean: what happened between the progenote and the cenancestor? Microbiol. SEM *11*, 185–198.

74. Lazcano A, F.G., and Oró, J. (1992). Life before DNA, the origin and evolution of early Archaean cells. In The Evolution of Metabolic Function, M. RP, ed. (CRC Press)), pp. 237–295.

75. Fournier, G.P., Andam, C.P., Alm, E.J., and Gogarten, J.P. (2011). Molecular evolution of aminoacyl tRNA synthetase proteins in the early history of life. Orig. Life Evol. Biosph. *41*, 621–632. https://doi.org/10.1007/s11084-011-9261-2.

76. Di Giulio, M. (2008). An extension of the coevolution theory of the origin of the genetic code. Biol. Direct *3*, 37. https://doi.org/10.1186/1745-6150-3-37.

77. Wong, J.T. (1975). A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA *72*, 1909–1912. https://doi.org/10.1073/pnas.72.5.1909.

78. Wong, J.T.F. (2005). Coevolution theory of the genetic code at age thirty. Bioessays *27*, 416–425. https://doi.org/10.1002/bies.20208.

79. De Tarafder, A., Parajuli, N.P., Majumdar, S., Kaçar, B., and Sanyal, S. (2021). Kinetic Analysis Suggests Evolution of Ribosome Specificity in Modern Elongation Factor-Tus from "Generalist" Ancestors. Mol. Biol. Evol. *38*, 3436–3444. https://doi.org/10.1093/molbev/msab114.

80. Wehbi, S., Wheeler, A., Morel, B., Manepalli, N., Minh, B.Q., Lauretta, D.S., and Masel, J. (2024). Order of amino acid recruitment into the genetic code resolved by last universal common ancestor's protein domains. Proc. Natl. Acad. Sci. USA *121*, e2410311121. https://doi.org/10.1073/pnas.2410311121.

81. Zwickl, D.J., and Hillis, D.M. (2002). Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. *51*, 588–598. https://doi.org/10.1080/10635150290102339.

82. Bernabeu, M., Manzano-Morales, S., and Gabaldón, T. (2024). On the impact of incomplete taxon sampling on the relative timing of gene transfer events. PLoS Biol. *22*, e3002460. https://doi.org/10.1371/journal.pbio.3002460.

83. Kolaczkowski, B., and Thornton, J.W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol. Biol. Evol. *25*, 1054–1066. https://doi.org/10.1093/molbev/msn042.

84. Wang, H.C., Minh, B.Q., Susko, E., and Roger, A.J. (2018). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst. Biol. *67*, 216–235. https://doi.org/10.1093/sysbio/syx068.

85. Chothia, C., and Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J. *5*, 823–826. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x.

86. Puente-Lelievre, C., Malik, A., and Douglas, J. (2025). Protein Structural Phylogenetics. Genome Biol. Evol. *17*, evaf139. https://doi.org/10.1093/gbe/evaf139.

87. Cuevas Zuviria, B., Detemple, F., Amritkar, K., Garcia, A.K., Seefeldt, L., Einsle, O., and Kacar, B. (2025). Structural evolution of nitrogenase over 3 billion years. eLife *14*, 105613. https://doi.org/10.7554/eLife.105613.

88. Harris, A.J., and Goldman, A.D. (2018). Phylogenetic Reconstruction Shows Independent Evolutionary Origins of Mitochondrial Transcription Factors from an Ancient Family of RNA Methyltransferase Proteins. J. Mol. Evol. *86*, 277–282. https://doi.org/10.1007/s00239-018-9842-z.

89. Fournier, G.P., Andam, C.P., and Gogarten, J.P. (2015). Ancient horizontal gene transfer and the last common ancestors. BMC Evol. Biol. *15*, 70. https://doi.org/10.1186/s12862-015-0350-0.

90. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

91. Nalecz-Charkiewicz, K., Charkiewicz, K., and Nowak, R.M. (2024). Quantum computing in bioinformatics: a systematic review mapping. Briefings Bioinf. *25*, bbae391. https://doi.org/10.1093/bib/bbae391.

92. Pyzer-Knapp, E.O., and Curioni, A. (2024). Advancing biomolecular simulation through exascale HPC, AI and quantum computing. Curr. Opin. Struct. Biol. *87*, 102826. https://doi.org/10.1016/j.sbi.2024.102826.

93. Kacar, B. (2024). Reconstructing Early Microbial Life. Annu. Rev. Microbiol. *78*, 463–492. https://doi.org/10.1146/annurev-micro-041522-103400.